Improving power posterior estimation of statistical evidence

Nial Friel*, Merrilee Hurn† and Jason Wyse‡

September 17, 2012

Abstract

The statistical evidence (or marginal likelihood) is a key quantity in Bayesian statistics, allowing one to assess the probability of the data given the model under investigation. This paper focuses on refining the power posterior approach to improve estimation of the evidence. The power posterior method involves transitioning from the prior to the posterior by powering the likelihood by a temperature variable. In common with other tempering algorithms, the power posterior involves some degree of tuning, and this paper addresses this issue. The main contributions of this article are twofold – we present a result from the numerical analysis literature which can reduce the bias in the estimate of the evidence by addressing the error arising from numerically integrating across the temperature. We also address the choice of temperature ladder, and present an adaptive algorithm which gives excellent performance in the examples considered here. A key practical point is that both of these innovations incur virtually no extra cost.

Keywords: Marginal likelihood, Markov chain Monte Carlo, Power posteriors, Statistical evidence, Tempering, Thermodynamic integration.

1 Introduction

The statistical evidence (sometimes called the marginal likelihood or integrated likelihood) is a vital quantity in Bayesian statistics for the comparison of models, $m_1, \ldots, m_l$. Under the Bayesian paradigm we consider the posterior distribution

$$p(\theta_i, m_i|y) \propto p(y|\theta_i, m_i)p(\theta_i|m_i)p(m_i), \quad \text{for} \quad i = 1, \ldots, l, \tag{1}$$

for data y and parameters θ_i within model m_i, where $p(\theta_i|m_i)$ denotes the prior distribution for parameters within model m_i and where $p(m_i)$ denotes the prior model probability. The evidence for data y given model

*School of Mathematical Sciences, University College Dublin, Belfield, Dublin 4, Republic of Ireland; Email: nial.friel@ucd.ie

†Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, UK; Email: M.A.Hurn@bath.ac.uk

‡School of Computer Science and Statistics, Trinity College Dublin, College Green, Dublin 2, Republic of Ireland; Email: wyseja@scss.tcd.ie

m_i arises as the normalising constant of the posterior distribution within model m_i,

$$p(\theta_i|y, m_i) \propto p(y|\theta_i, m_i)p(\theta_i|m_i), \tag{2}$$

and thus results from integrating the un-normalised posterior across the θ_i parameter space,

$$p(y|m_i) = \int_{\theta_i} p(y|\theta_i, m_i)p(\theta_i|m_i)\, d\theta_i. \tag{3}$$

This of course assumes that the prior distribution for θ_i is proper. The marginal likelihood is often then used to calculate Bayes factors when one wants to compare two competing models, m_i and m_j,

$$BF_{ij} = \frac{p(y|m_i)}{p(y|m_j)} = \frac{p(m_i|y)}{p(m_j|y)}\frac{p(m_j)}{p(m_i)}. \tag{4}$$

Here, $p(m_i|y)$ is the posterior probability for model m_i and it can be evaluated, using the evidence for each of the collection of models under consideration,

$$p(m_i|y) \propto p(y|m_i)p(m_i), \quad \text{for } \; i = 1, \ldots, l. \tag{5}$$

Estimation of the evidence is a non-trivial task for most statistical models and there has been considerable effort in the literature to find algorithms and methods for this purpose. Laplace's method (Tierney and Kadane, 1986) is an early approach and very widely used. Other notable and popular approaches include Chib's method (Chib 1995), annealed importance sampling (Neal 2001), nested sampling (Skilling 2006), bridge sampling (Meng and Wong, 1996) and power posteriors (Friel and Pettitt, 1998) which is the focus of this paper. For a recent review and perspective on these and other methods, see Friel and Wyse (2012).

This paper is organised as follows. Section 2 outlines the power posterior method, and the approach we propose to improve estimation of the evidence. Section 3 illustrates the potential gain from implementing the methodology which we propose. We offer some conclusions in Section 4.

2 The power posterior approach

In what follows we will drop the explicit conditioning on model m_i for notational simplicity. We follow the notation of Friel and Pettitt (2008) and denote the power posterior by

$$p_t(\theta|y) \;\propto\; p(y|\theta)^t p(\theta),\; t \in [0, 1] \tag{6}$$

$$\text{with } z(y|t) \;=\; \int_\theta p(y|\theta)^t p(\theta) d\theta. \tag{7}$$

where $t \in [0, 1]$ is thought of as a temperature, which has the effect of tempering the likelihood, whereby at the extreme ends of the temperature range, $p_0(\theta|y)$ and $p_1(\theta|y)$ correspond to the prior and posterior,

respectively. The power posterior estimator for the evidence relies on noting that

$$\begin{aligned}
\frac{d}{dt}\log(z(y|t)) &= \frac{1}{z(y|t)}\frac{d}{dt}z(y|t) \\
&= \frac{1}{z(y|t)}\int_\theta \frac{d}{dt}p(y|\theta)^t p(\theta)d\theta \\
&= \frac{1}{z(y|t)}\int_\theta p(y|\theta)^t \log(p(y|\theta))p(\theta)d\theta \\
&= \int_\theta \frac{p(y|\theta)^t p(\theta)}{z(y|t)}\log(p(y|\theta))d\theta \\
&= \mathbf{E}_{\theta|y,t}\log(p(y|\theta)). \qquad (8)
\end{aligned}$$

As a result

$$\begin{aligned}
\int_0^1 \mathbf{E}_{\theta|y,t}\log(p(y|\theta))dt &= [\,\log(z(y|t))\,]_0^1 \\
&= \log(z(y|t=1)) \text{ (assuming that the prior is normalised)} \qquad (9)
\end{aligned}$$

which is the log of the desired marginal likelihood.

In practice the temperature range is discretised as $0 = t_0 < t_1, \ldots, t_n = 1$ to form an estimator based on (9). For each t_i, a sample from $p(\theta|y, t_i)$ can be used to estimate $\mathbf{E}_{\theta|y,t_i}\log(p(y|\theta))$. Finally, a trapezoidal rule is used to approximate

$$\log p(y) \approx \sum_{i=1}^{n}(t_i - t_{i-1})\left(\frac{\mathbf{E}_{\theta|y,t_{i-1}}\log(p(y|\theta)) + \mathbf{E}_{\theta|y,t_i}\log(p(y|\theta))}{2}\right). \qquad (10)$$

Discretising t introduces an approximation into this method and the two goals of this paper are to reduce the bias in the power posterior estimation method due to the approximation and also to find an adaptive method for choosing the temperature rungs required. For both of these we will exploit the fact that the gradient of the expected log deviance curve equals its variance, as we now outline.

Differentiating $\mathbf{E}_{\theta|y,t}\log(p(y|\theta))$ with respect to t yields

$$\begin{aligned}
\frac{d}{dt}\mathbf{E}_{\theta|y,t}\log(p(y|\theta)) &= \int_\theta \log(p(y|\theta))\frac{d}{dt}p_t(\theta|y)d\theta \\
&= \int_\theta \log(p(y|\theta))\left[\log(p(y|\theta)) - \frac{1}{z(y|t)}\frac{d}{dt}z(y|t)\right]p_t(\theta|y)d\theta \\
&= \int_\theta \log(p(y|\theta))\left[\log(p(y|\theta)) - \frac{d}{dt}\log(z(y|t))\right]p_t(\theta|y)d\theta \\
&= \mathbf{E}_{\theta|y,t}\log(p(y|\theta))^2 - (\mathbf{E}_{\theta|y,t}\log(p(y|\theta)))^2 \\
&= \mathbf{V}_{\theta|y,t}(\log(p(y|\theta))) \qquad (11)
\end{aligned}$$

where $\mathbf{V}_{\theta|y,t}(\log(p(y|\theta)))$ denotes the variance of the log deviance at temperature t.

2.1 Reducing the bias by improving the numerical integration

Equation (11) immediately provides two useful pieces of information. First, the curve which we wish to integrate numerically is (strictly) increasing. Secondly, we can improve upon the standard trapezium rule used to numerically integrate the expected log deviance by incorporating derivative information at virtually no extra computational cost (the cost merely of calculating the variance of a set of simulations for fixed t). We do this by using the corrected trapezium rule which comes from an error analysis of the standard trapezium rule, see for example Atkinson and Han (2004), Section 5.2; when integrating a function f between points a and b

$$\int_a^b f(x)dx = (b-a)\left[\frac{f(b)+f(a)}{2}\right] - \frac{(b-a)^3}{12}f''(c) \tag{12}$$

where c is some point in $[a,b]$. The first term of the right hand side of this equation is the usual trapezium rule and the second can be approximated using

$$\begin{aligned} f''(c) &\approx \frac{f'(b)-f'(a)}{b-a} \\ \text{so that } \int_a^b f(x)dx &\approx (b-a)\left[\frac{f(b)+f(a)}{2}\right] - \frac{(b-a)^2}{12}\left[f'(b)-f'(a)\right]. \end{aligned} \tag{13}$$

This latter form motivates the corrected trapezium rule which for unequally spaced x-axis points, taken together with the information derived above regarding the derivative of the log deviance gives

$$\begin{aligned} \log(z(y|t=1)) \approx & \sum_{i=0}^{n-1}(t_{i+1}-t_i)\left[\frac{\mathbf{E}_{\theta|y,t_i}\log(p(y|\theta)) + \mathbf{E}_{\theta|y,t_{i+1}}\log(p(y|\theta))}{2}\right] \\ & - \sum_{i=0}^{n-1}\frac{(t_{i+1}-t_i)^2}{12}\left[\mathbf{V}_{\theta|y,t_{i+1}}(\log(p(y|\theta))) - \mathbf{V}_{\theta|y,t_i}(\log(p(y|\theta)))\right] \end{aligned} \tag{14}$$

where both the expectations $\{\mathbf{E}_{\theta|y,t_i}\log(p(y|\theta))\}$ and variances $\{\mathbf{V}_{\theta|y,t_i}(\log(p(y|\theta)))\}$ are to be estimated using MCMC runs at a number of values of t_i.

2.2 Adaptive choice of the temperature placement

The next question which arises is how to choose the $\{t_i\}$ between $t_0 = 0$ and $t_n = 1$. Friel and Pettitt (2008) find that setting $t_i = (i/n)^5$ performs well. We refer to this as the powered fraction (PF) schedule. Lartillot and Philippe (2006) discuss very similar ideas in the phylogenetics literature, although using Simpson's rule for the numerical integration; they use equally spaced temperatures between 0 and 1.

Here we will only consider the discretisation error associated with the numerical integration, rather than the stochastic error arising with sampling from the different $p_{t_i}(\theta|y)$. Calderhead and Girolami (2009) show that this discretisation error depends upon the Kullback-Liebler distance between successive $p_{t_i}(\theta|y)$.

Lefebvre, Steele and Vandal (2010) also consider a symmetrised Kullback-Liebler divergence in picking optimal schedules for path sampling. At first glance the Kullback-Liebler distance does not seem a particularly tractable quantity to manipulate. However, these papers and Behrens, Friel and Hurn (2012) all note that, in the notation of this paper,

$$\sum_{i=1}^{n-1}(KL[p_{t_i}(\theta|y), p_{t_{i+1}}(\theta|y)] + KL[p_{t_{i+1}}(\theta|y), p_{t_i}(\theta|y)]) = 2S_n(t_0, \ldots, t_n) \tag{15}$$

where KL denotes the Kullback-Liebler distance and

$$S_n(t_0, \ldots, t_n) = \sum_{i=0}^{n-1}(t_{i+1} - t_i)\mathbf{E}_{\theta|y,t_{i+1}} \log(p(y|\theta)) - \sum_{i=0}^{n-1}(t_{i+1} - t_i)\mathbf{E}_{\theta|y,t_i} \log(p(y|\theta)). \tag{16}$$

S_n can be interpreted graphically as the sum of the rectangular areas between a lower and an upper approximation to the integral of $\mathbf{E}_{\theta|y,t_i} \log(p(y|\theta))$ between $t_0 = 0$ and $t_1 = 1$. Behrens, Friel and Hurn (2012) use minimising S_n as a rationale for choosing the temperatures in tempered transitions. We propose to use the same target in selecting the $\{t_i\}$ for power posteriors. However, unlike in tempered transitions where the tuning forms a small part of the overall computational load, here the cost is almost exclusively the estimation of $\mathbf{E}_{\theta|y,t_i} \log(p(y|\theta))$. We propose the following scheme: Initialise a set of m $\{t_i\}$ using the geometric placing including 0 and 1 (we will see in later examples why a reasonable starting point is necessary) where m is a small proportion of the proposed total number of rungs n. These m $\{t_i\}$ contribute $(m-1)$ terms $\{[t_{i+1} - t_i][\mathbf{E}_{\theta|y,t_{i+1}} \log(p(y|\theta)) - \mathbf{E}_{\theta|y,t_i} \log(p(y|\theta))]\}$ which sum to give S_n. Identify the largest of these terms and locate the next point in the corresponding interval, say $[t_k, t_{k+1}]$. Since we do not want to use computational resources in performing a search for the optimal location of the new t_i (there is no analytic solution), we follow a low cost route using the estimated gradients/variances at t_k and t_{k+1}. If the estimated gradient at t_k is denoted by $\hat{V}_k$ and that at t_{k+1} by $\hat{V}_{k+1}$, we set the new point to be

$$t = t_k + \frac{\hat{V}_{k+1}}{\hat{V}_k + \hat{V}_{k+1}}(t_{k+1} - t_k). \tag{17}$$

This scheme will almost certainly not identify the optimal placing of the n rungs. However it is quick, cheap and intuitively reasonable. (In practice, Monte Carlo error can mean that the function is not increasing and so the criterion is changed to picking the interval with the largest absolute contribution to S_n.)

3 Examples

We present three examples which illustrate the gains that arise from employing the methods developed here. The first example is a non-nested linear regression comparison for which the marginal likelihoods can be calculated analytically. Example 2 is a larger problem, choosing between two logistic regression models, for

which an analytic solution is not possible. These first two examples were included in the review paper by Friel and Wyse (2012) where the performance of power posteriors was compared to other existing methods. The final example is by far the largest and exhibits the most interestingly shaped $\mathbf{E}_{\theta|y,t}\log(p(y|\theta))$.

3.1 Example 1: Radiata pine

The first example compares two linear regression models for the Radiata pine data originally in Williams (1959). The response variable here is the maximum compression strength parallel to the grain, y_i, while the predictors are density, x_i, or density adjusted for resin content, z_i, for $n = 42$ specimens of radiata pine. Two possible Gaussian linear regression models are considered;

$$\begin{aligned}
&\text{Model 1:} \quad y_i = \alpha + \beta(x_i - \bar{x}) + \epsilon_i, \quad \epsilon_i \sim N(0, \tau^{-1}), \quad i = 1, \ldots, n, \\
&\text{Model 2:} \quad y_i = \gamma + \delta(z_i - \bar{z}) + \eta_i, \quad \eta_i \sim N(0, \lambda^{-1}), \quad i = 1, \ldots, n.
\end{aligned}$$

Priors are chosen to match the analyses of Friel and Wyse (2012) (baring a notational factor of 2). The regression parameters $(\alpha, \beta)^T$ and $(\gamma, \delta)^T$ are taken to be Normally distributed with mean $(3000, 185)^T$ and precision τQ_0 and λQ_0 respectively where $Q_0 = \text{diag}(r_0, s_0)$. The values of r_0 and s_0 were fixed to be 0.06 and 6. A gamma prior with shape $a_0 = 3$ and rate $b_0 = 2 \times 300^2$ was assumed for both τ and λ.

Following the comparisons of Friel and Wyse (2012), we consider estimating the evidence using 10, 20, 50, 100 or 200 rungs in the tempering scheme. The parameters at all levels are updated using the Gibbs sampler. For this example, both the adaptive and the PF spacings use 20000 iterations at each rung, discarding the first fifth of these as burn in. Figure 1 shows the expected log deviance curves for the two models using 200 rungs, their shapes suggesting that PF spacing might perform competitively (Behrens, Friel and Hurn (2012) show that a scheme where t_i/t_{i+1} is a constant for $i > 0$ minimises S_n when the integrand takes the form $\frac{K_1}{t} + K_2$ for some constants K_1 and K_2).

Figure 2 shows the upper and lower bounds of the evidence (in black), the uncorrected estimate (in red) and the corrected estimate (in blue) all for model 1 as the number of rungs increases. The PF spacing results are denoted by solid lines and the adaptive spacing results by dashed lines. The true value of the evidence is known for this example and is marked by a horizontal line. As the vertical scale differs quite significantly between $n = 10$ and $n = 200$, the figure is split into two plots, small numbers of rungs (where the upper and lower bounds are not tight) and large numbers of rungs. The adaptive temperature placement is initialised using the 10 rung PF placement. Since for the adaptive spacing, increasing the number of rungs by one requires only one additional set of MCMC iterations at the new temperature, there is an averaging effect and the dashed lines appear smoother than the solid ones (where for an increase of one rung, all temperatures apart from $t_0 = 0$ and $t_n = 1$ change and so the estimates at successive rungs are independent

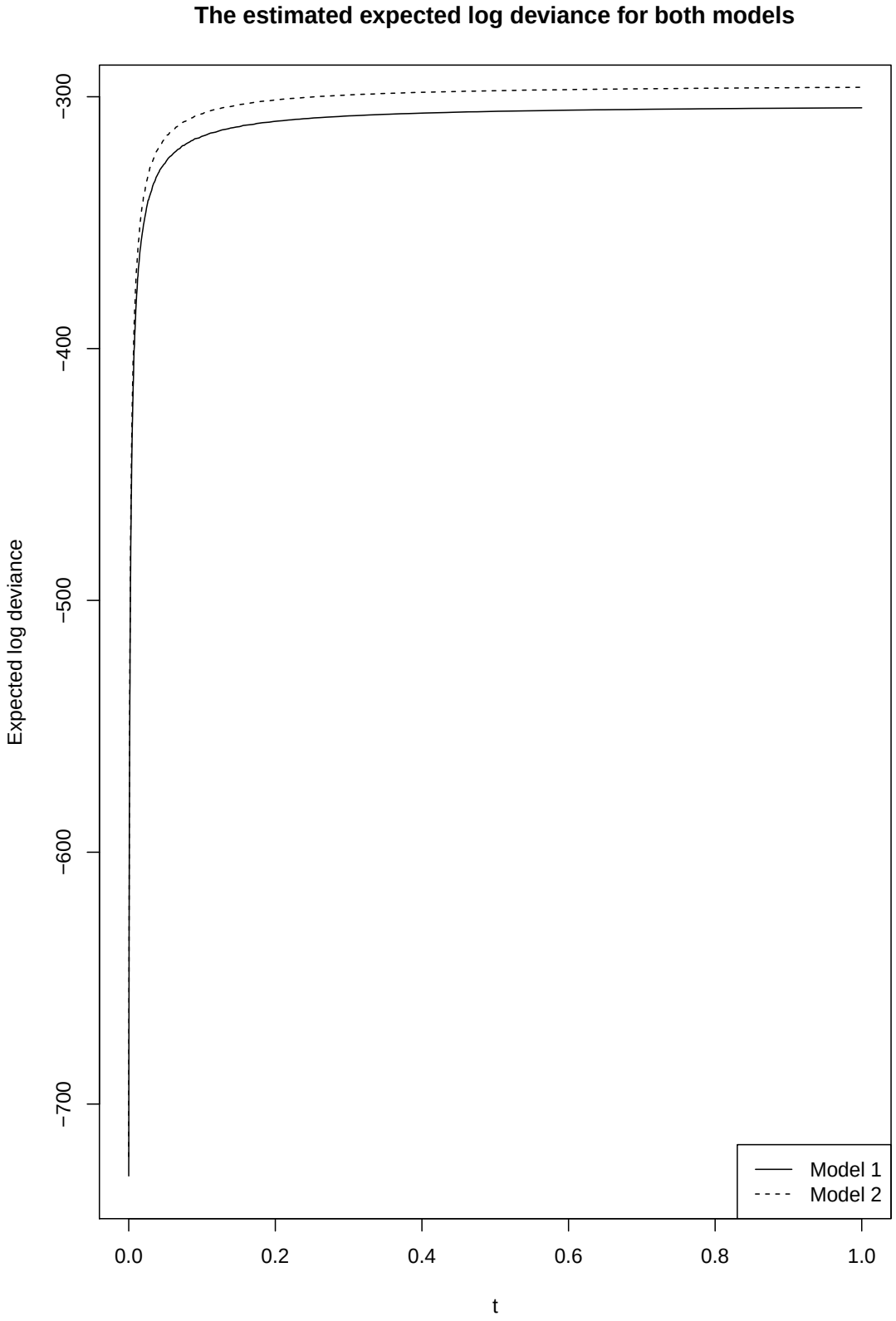
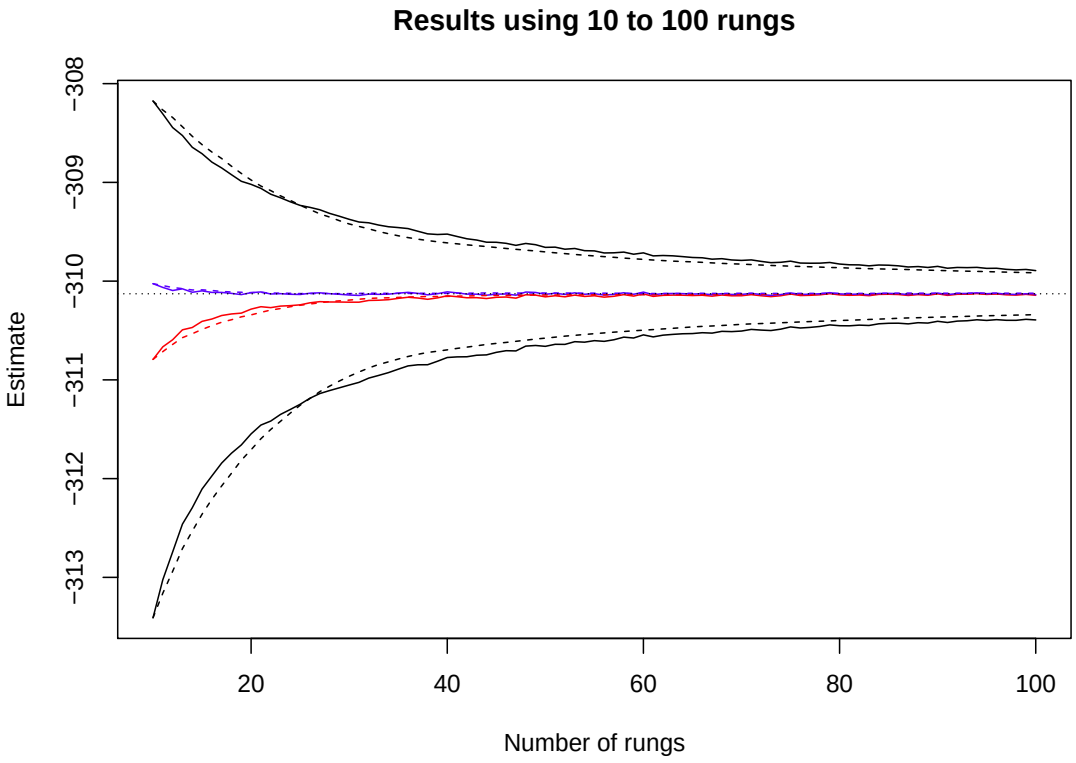


Figure 1: The expected log deviance curves for the two Radiata models using 200 rungs.



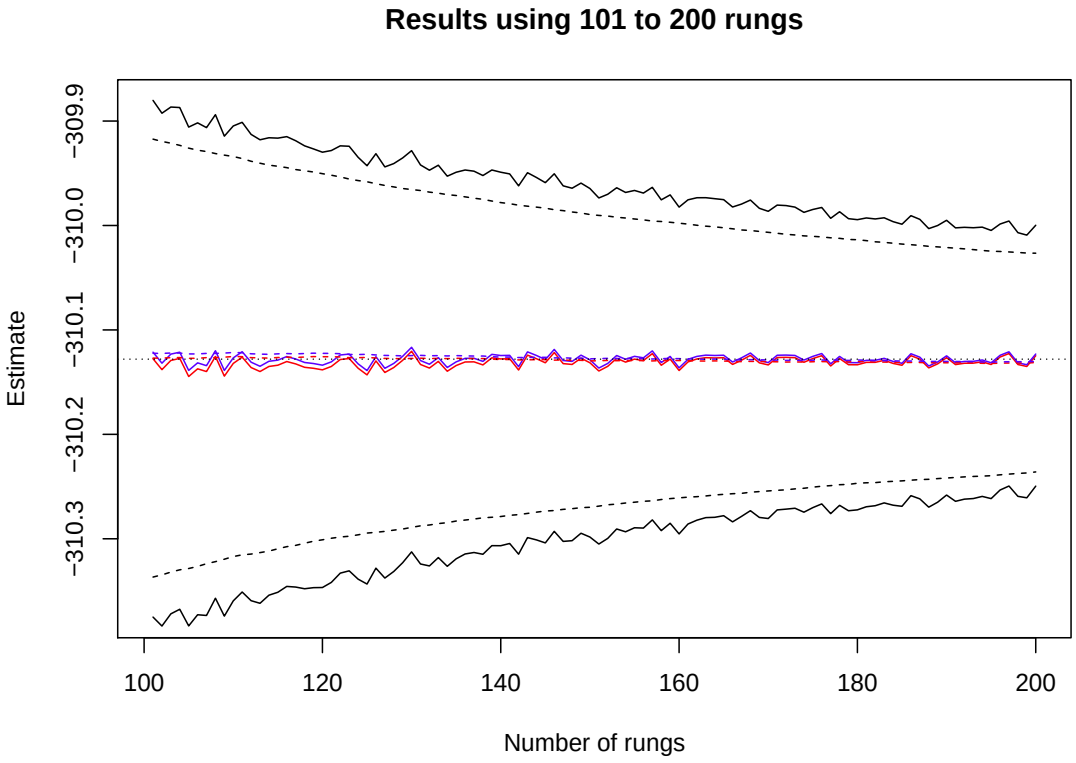


Figure 2: For the Radiata example. Upper and lower bounds (in black), uncorrected estimates (red), corrected estimates (blue) as the number of rungs increases for model 1. Solid lines indicate PF spacing, dashed lines the adaptive schedule.

		10 rungs	20 rungs	50 rungs	100 rungs	200 rungs
Model 1	PF uncorrected	-0.6493	-0.1607	-0.0260	-0.0060	-0.0030
		(0.0271)	(0.0175)	(0.0098)	(0.0081)	(0.0056)
	PF corrected	0.1042	0.0066	-0.0002	0.0005	-0.0014
		(0.0211)	(0.0166)	(0.0097)	(0.0080)	(0.0056)
	Adaptive uncorrected	-0.6543	-0.2137	-0.0211	-0.0041	-0.0008
		(0.0223)	(0.0142)	(0.0090)	(0.0061)	(0.0053)
	Adaptive corrected	0.0995	0.0130	-0.0009	0.0006	0.0003
		(0.0175)	(0.0129)	(0.0090)	(0.0061)	(0.0053)
Model 2	PF uncorrected	-0.6375	-0.1514	-0.0232	-0.0049	-0.0008
		(0.0279)	(0.0176)	(0.0106)	(0.0074)	(0.0057)
	PF corrected	0.0990	0.0108	0.0019	0.0013	0.0008
		(0.0215)	(0.0166)	(0.0105)	(0.0073)	(0.0057)
	Adaptive uncorrected	-0.6395	-0.2112	-0.0193	-0.0038	-0.0005
		(0.0306)	(0.0207)	(0.0094)	(0.0077)	(0.0042)
	Adaptive corrected	0.0987	0.0104	0.0002	0.0007	0.0006
		(0.0248)	(0.0168)	(0.0093)	(0.0077)	(0.0042)

Table 1: Estimated bias (and standard deviation) in estimating the evidence for the two Radiata pine models.

of one another). From this figure, it appears that the corrected estimates converge faster towards the true value than do the uncorrected ones initially. By construction, the adaptive and PF schedules coincide at 10 rungs. Immediately after that, the adaptive schedule initially provides wider bounds on the evidence but after approximately 25 rungs, the bounds are consistently narrower.

To quantify these observations, the bias is estimated as per the approach of Friel and Wyse (2012), performing 50 replicates at 10, 20, 50, 100 and 200 rungs using 10000 iterations of which the first fifth are discarded as burn in. The average and standard deviation of the 50 biases are given in Table 1. As might be expected from the concave shape of the log deviance curves, the uncorrected integrals tend to underestimate the evidence, giving negative biases which decrease as the number of rungs increases. To visualise the effect of both the correction and the adaptive placing of temperatures, Figure 3 plots the 50 observed biases separately for each number of rungs and under the two spacings with or without correction. Correction is particularly effective when smaller numbers of rungs are being used. The final panel illustrates the effect of the correction when using 100 rungs by plotting the corrected points against their uncorrected values (the

line $y = x$ is shown in red, corresponding to no effect of correction). There is an interesting difference between the adaptive and the PF versions, less correction is needed for the adaptive schedule, that is, it is doing a better job of the numerical integration.

Given the good reductions in bias seen in Table 1, it is important to ask how much extra time is required. To assess this, a total of 10 runs for model 1 using 20000 iterations and 200 temperatures were timed. Four versions of the algorithm were considered, corresponding to Table 1. All the coding was in R and times are given relative to the PF non-corrected version:

PF non-corrected	PF corrected	Adaptive non-corrected	Adaptive corrected
1.0000	1.0054	1.0017	0.9986

The adaptive selection of temperatures and the correction term in the numerical integration come at negligible computational cost. Given the reductions in bias achievable by the correction in particular, there is no reason at all not to adopt this modification.

3.2 Example 2: Pima indians

We turn next to the Pima Indian example considered by Friel and Wyse (2012), originally described by Smith *et al* (1988) . These data record diabetes incidence and possible disease indicators for $n = 532$ Pima Indian women aged over 20. The seven possible disease indicators are the number of pregnancies (`NP`), plasma glucose concentration (`PGC`), diastolic blood pressure (`BP`), triceps skin fold thickness (`TST`), body mass index (`BMI`), diabetes pedigree function (`DP`) and age (`AGE`), with all these covariates standardised.

The model assumed for the observed diabetes incidence, $y = (y_1, \ldots, y_n)$, is

$$p(y|\theta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \tag{18}$$

where p_i is the probability of incidence for person i, and p_i is related to the i^{th} person's covariates and a constant term, denoted by $x_i = (1, x_{i1}, \ldots, x_{id})^T$, and the parameters, $\theta = (\theta_0, \theta_1, \ldots, \theta_d)^T$, by

$$\log\left(\frac{p_i}{1-p_i}\right) = \theta^T x_i \tag{19}$$

where d is the number of explanatory variables. An independent multivariate Gaussian prior is assumed for θ, with mean zero and non-informative precision of $\tau = 0.01$, so that

$$p(\theta) = (2\pi)^{-d/2}\tau^{d/2}\exp\left\{-\frac{\tau}{2}\theta^T\theta\right\}. \tag{20}$$

There are 129 potential models (2^7 models with covariates plus a model with only a constant term). A long reversible jump run (Green, 1995) revealed the two models with the highest posterior probability:

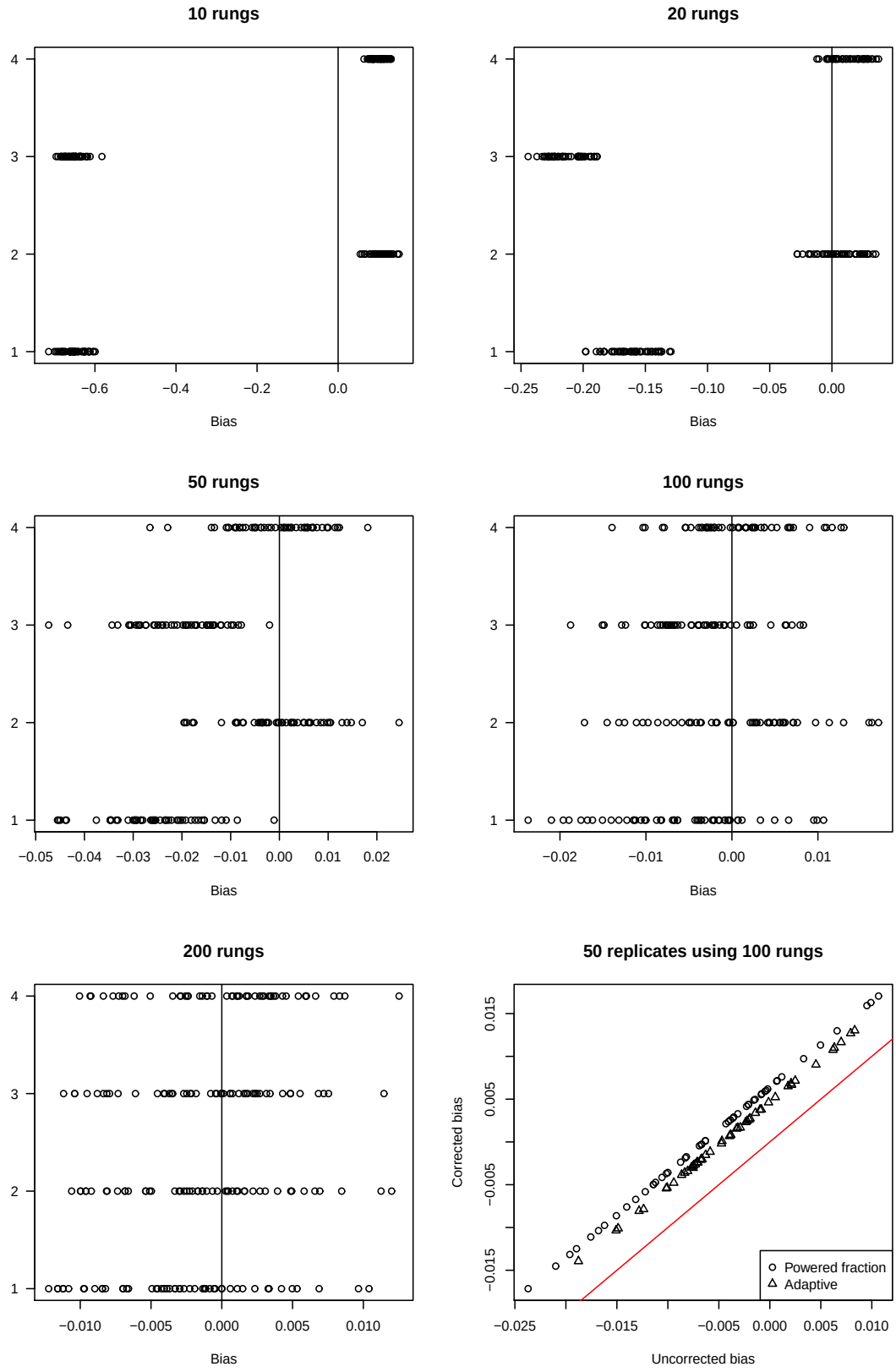


Figure 3: The 50 observed biases for model 1 using different numbers of rungs and the four schemes for the Radiata data: 1=Uncorrected PF spacing; 2=Corrected PF spacing; 3=Uncorrected adaptive spacing; 4=Corrected adaptive spacing. The red line in the comparison plot corresponds to the correction having no effect.

Model 1: `logit(p)= 1+NP+PGC+BMI+DP`

Model 2: `logit(p)= 1+NP+PGC+BMI+DP+AGE`

Figure 4 shows the estimated log deviance curves for these two competing models. Although the shapes are roughly similar to those in the Radiata example, notice the difference in scale on the y-axis compared to Figure 1. For these two models, the power posterior is not amenable to the Gibbs sampler and so a Metropolis update is used instead. This raises the problem of proposal scaling at the different temperatures. Since both the correction and the adaptive temperature placements rely on having good estimates of the variance of the log deviance, mixing is an important issue here. As an alternative to the approach taken in Friel and Wyse (2012), we have considered here a joint update of all the model parameters using a multivariate Normal proposal centred at the current value and with diagonal variance matrix, entries $\min(0.01/t, 1/\tau)$ where t is the temperature and τ is the precision of the prior. Run lengths are doubled compared to the previous example, using 40000 iterations and discarding the first fifth as burn in, with the adaptive temperature placement still initialised at the 10 rung PF placement.

Figure 5 shows the comparison of a single adaptive run with the corresponding PF scheme. Unlike in the Radiata example where the evidence can be evaluated analytically, Friel and Wyse (2012) use the Laplace approximation of the log evidence (Tierney and Kadane, 1986) as the "benchmark" in assessing convergence and bias. However this is not necessarily very accurate and so we replace the Laplace approximation by a very long run (2000 rungs) of the power posterior approach; the estimates we get are -257.2342 and -259.8519 for models 1 and 2 respectively as opposed to the Laplace approximations of -257.2588 and -259.8906. Comparing Figures 2 and 5, the greater y-axis scale of the expected log deviance curve for this example has led, not surprisingly, to wider intervals for the evidence. As before though, the adaptive temperature scheme initially under-performs the PF scheme but as the number of rungs increases beyond about 25 it generates narrower upper and lower bounds. The corrected estimates also appear to converge faster than the uncorrected ones towards the benchmark log evidence.

Table 2 shows the estimated biases and standard deviations (here estimated using 25 replicates rather than the 50 of the previous example for reasons of speed). These estimated biases are also depicted in Figure 6. Both the correction and the adaptation seem effective in reducing bias, the former particularly dramatically for small numbers of rungs while for the latter there is the same "catch-up" effect comparing the uncorrected versions for small numbers of rungs but thereafter there is a clear benefit. Interestingly, as in the previous example, the correction evens out the differences between the PF and the adaptive schedules' biases, although as the last panel in Figure 6 shows, to do so requires a larger correction for the PF.

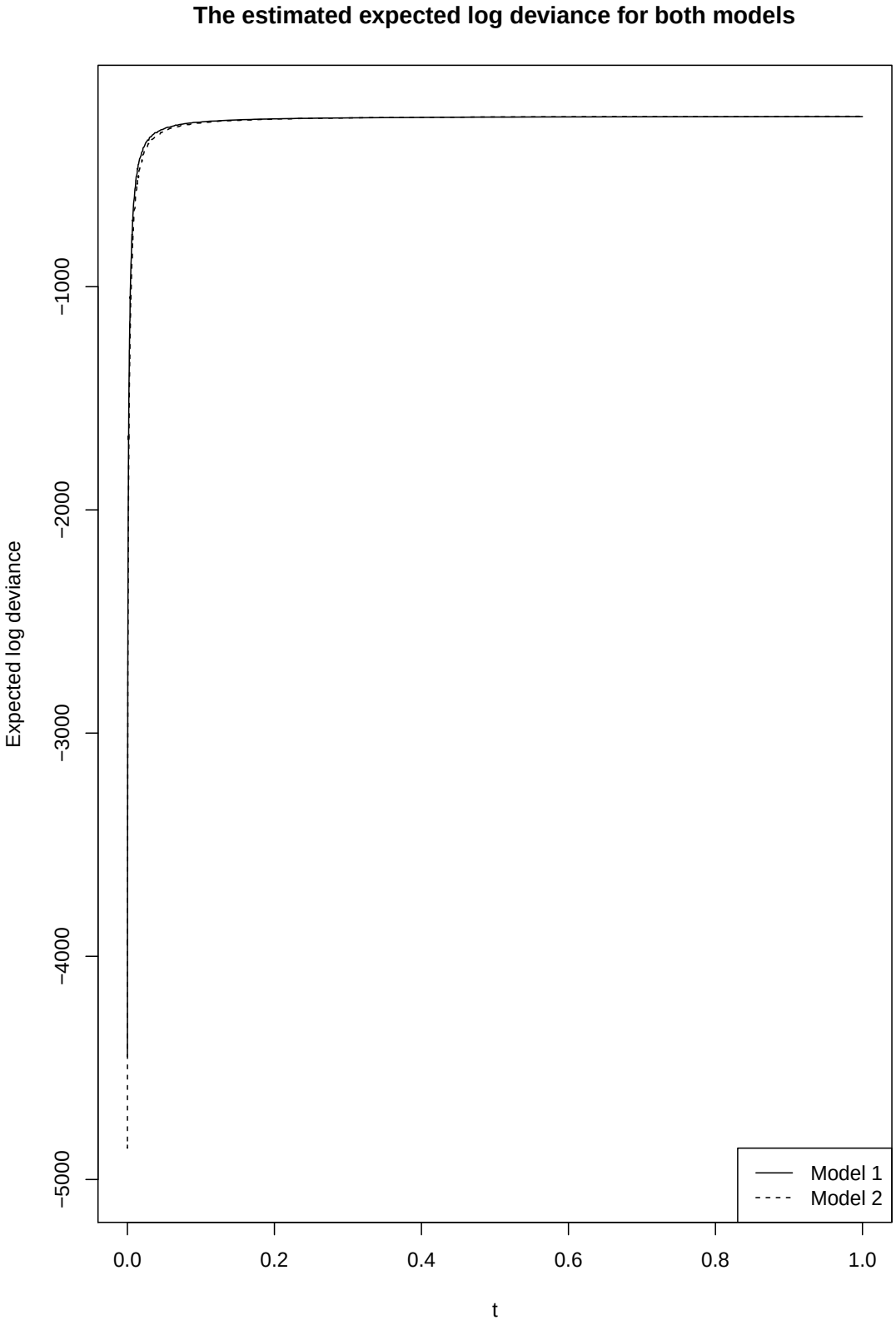
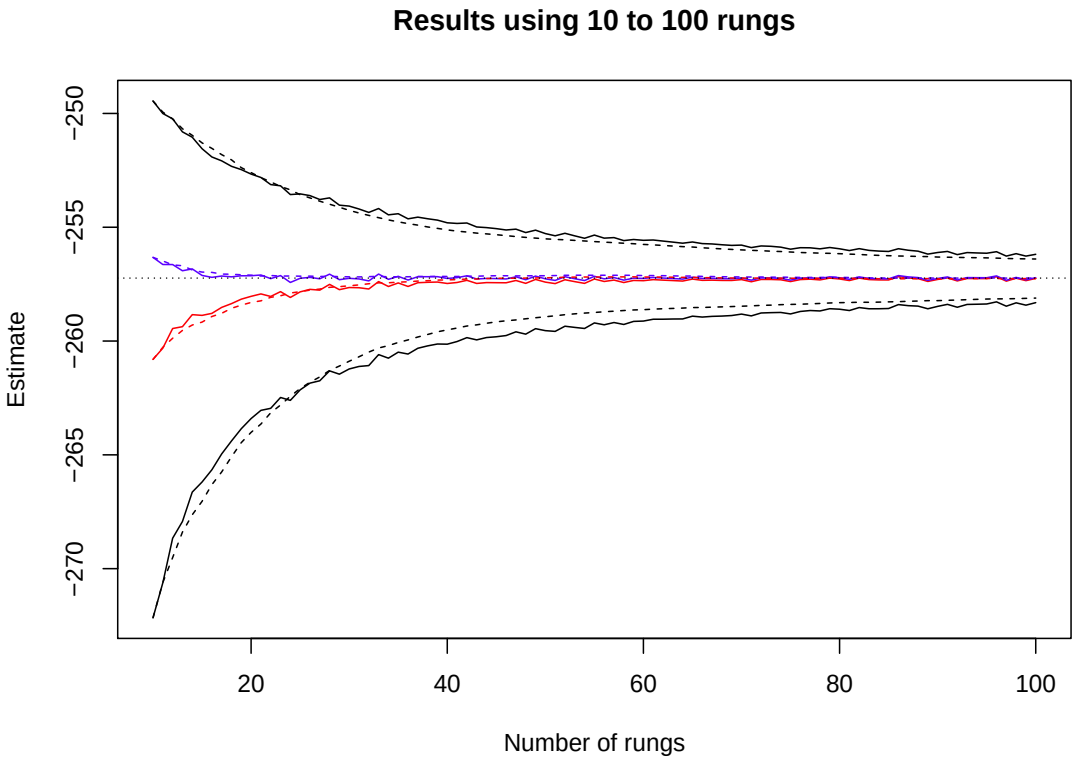


Figure 4: The expected log deviance curves for the two Pima models using 200 rungs.



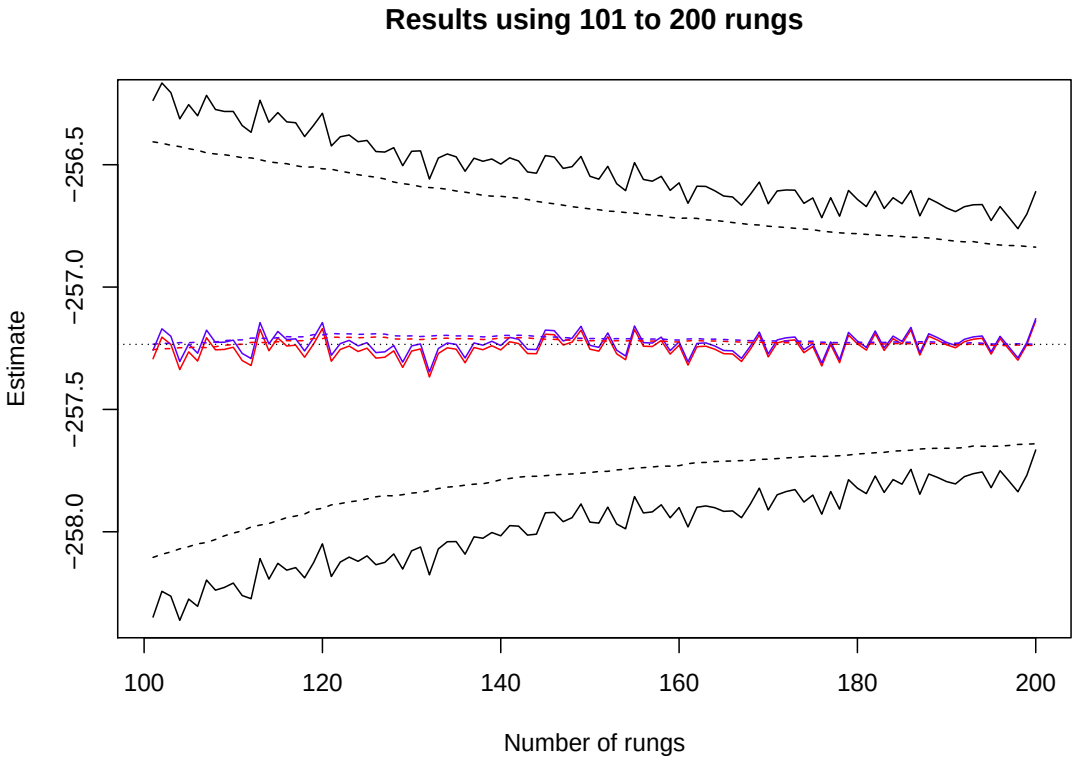


Figure 5: For the Pima example. Upper and lower bounds (in black), uncorrected estimates (red), corrected estimates (blue) as the number of rungs increases for model 1. Solid lines indicate PF spacing, dashed lines the adaptive schedule.

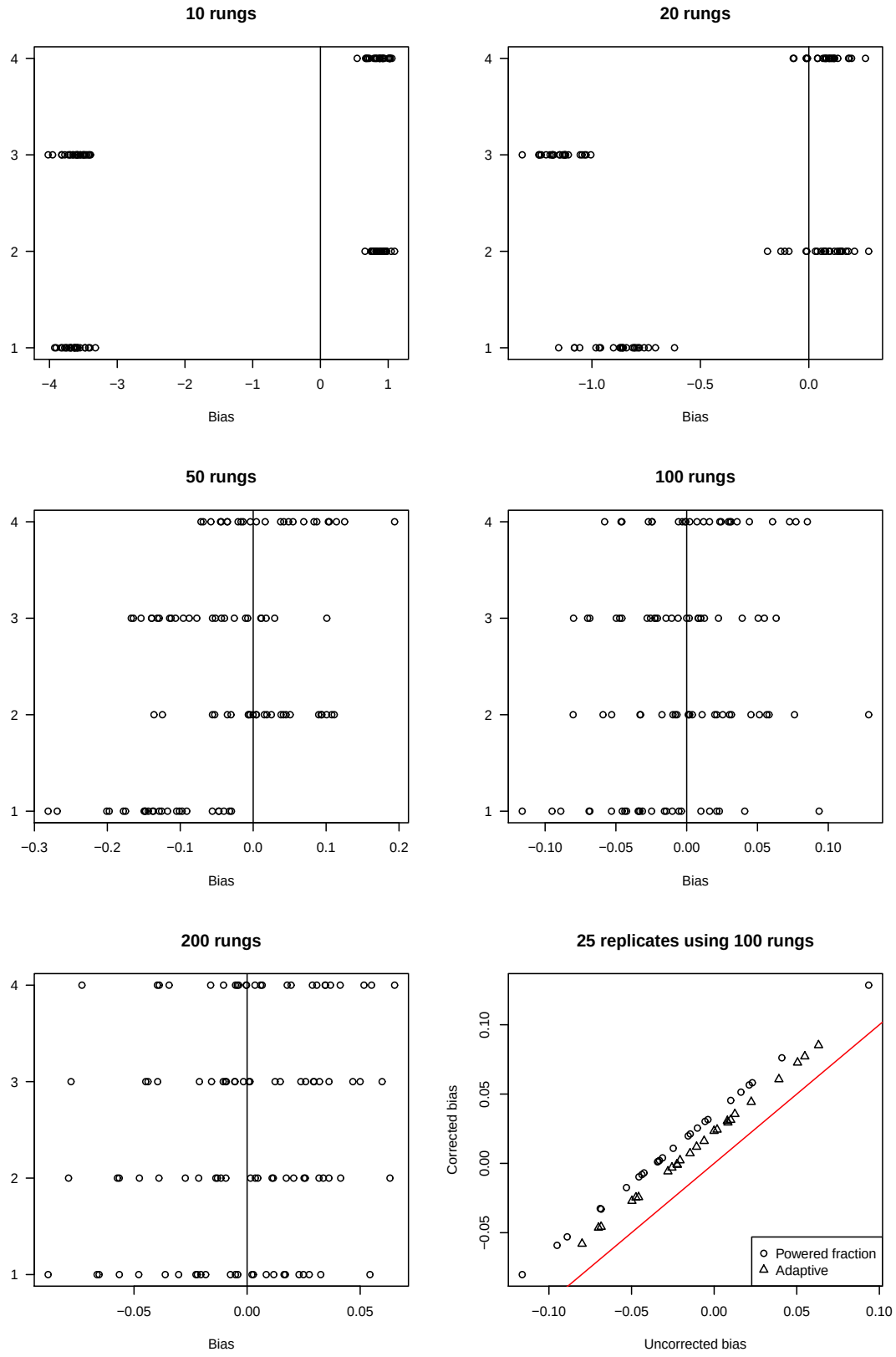


Figure 6: The 25 observed biases for model 1 using different numbers of rungs and the four schemes for the Pima Indian data: 1=Uncorrected PF spacing; 2=Corrected PF spacing; 3=Uncorrected adaptive spacing; 4=Corrected adaptive spacing.The red line in the comparison plot corresponds to the correction having no effect.

		10 rungs	20 rungs	50 rungs	100 rungs	200 rungs
Model 1	PF uncorrected	-3.6417	-0.8722	-0.1274	-0.0250	-0.0108
		(0.1509)	(0.1270)	(0.0671)	(0.0458)	(0.0352)
	PF corrected	0.8764	0.0699	0.0155	0.0106	-0.0019
		(0.1041)	(0.1117)	(0.0661)	(0.0455)	(0.0352)
	Adaptive uncorrected	-3.6202	-1.1496	-0.0673	-0.0096	0.0031
		(0.1640)	(0.0782)	(0.0706)	(0.0386)	(0.0331)
	Adaptive corrected	0.8656	0.0838	0.0268	0.0126	0.0082
		(0.1398)	(0.0791)	(0.0703)	(0.0384)	(0.0331)
Model 2	PF uncorrected	-4.1477	-1.0254	-0.1643	-0.0433	0.0140
		(0.2028)	(0.1172)	(0.0726)	(0.0359)	(0.0466)
	PF corrected	0.9633	0.0538	0.0004	-0.0025	-0.0039
		(0.1697)	(0.1074)	(0.0715)	(0.0357)	(0.0465)
	Adaptive uncorrected	-4.1619	-1.3083	-0.1165	-0.0281	-0.0020
		(0.1606)	(0.1025)	(0.0521)	(0.0458)	(0.0316)
	Adaptive corrected	1.0125	0.1045	-0.0057	-0.0020	0.0041
		(0.1504)	(0.0980)	(0.0515)	(0.0457)	(0.0316)

Table 2: Estimated bias (and standard deviation) in estimating the evidence for the two Pima Indian models.

3.3 Example 3: Galaxy data

To demonstrate a large application with a more challenging integral than the previous two, we use the much-studied Galaxy data set, see for example Richardson and Green (1997), which comprises measurements on the velocities of 82 galaxies. Denoting the 82 measurements by $y = \{y_1, \ldots, y_{82}\}$, we follow Richardson and Green (1997) in incorporating corresponding latent allocation variables $z = \{z_1, \ldots, z_{82}\}$. Given $z_i = j$, y_i follows the j^{th} of the k component Gaussian distributions of the mixture,

$$p(y_i|z_i = j, \mu_j, \sigma_j^2) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(\frac{-(y_i - \mu_j)^2}{2\sigma_j^2}\right) \quad i = 1, \ldots, 82. \tag{21}$$

Conditional independence is assumed for the $\{y_i\}$ and we specify independent standard proper priors:

$$\mathsf{P}(z_i = j) = w_j, \quad \text{where } \sum_{j=1}^{k} w_j = 1 \tag{22}$$

$$\{w_1, \ldots, w_k\} \sim Dirichlet(1, k) \tag{23}$$

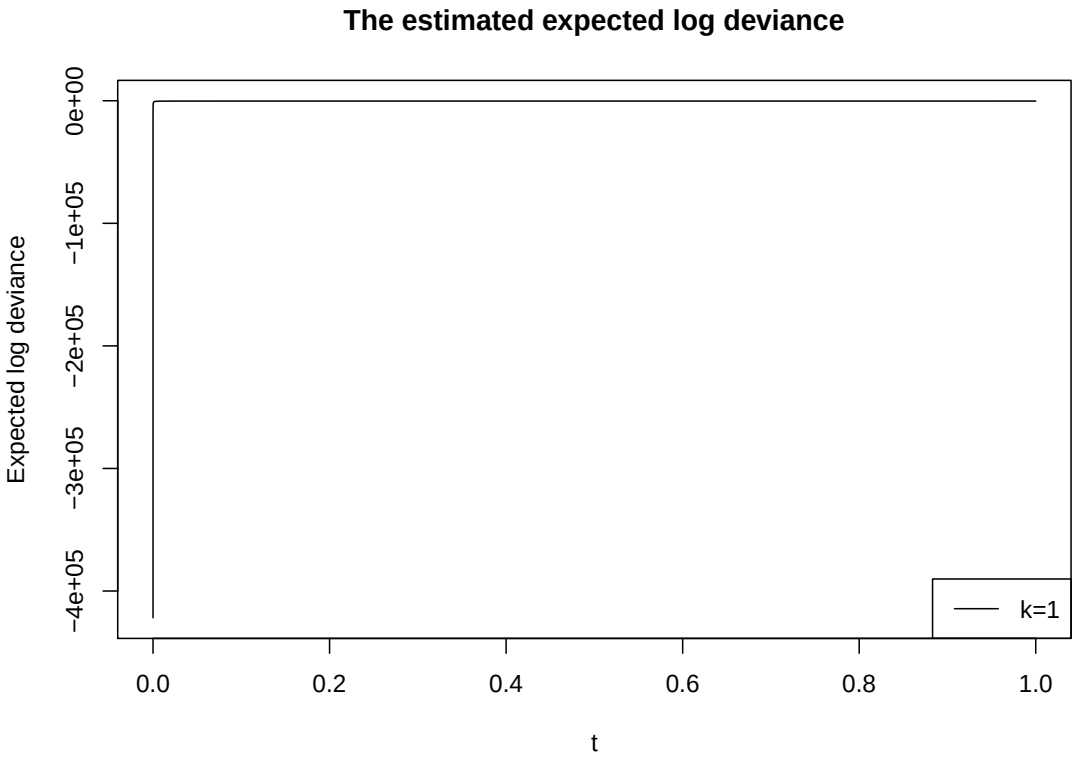
$$\mu_j \sim N(0, 1000), \quad j = 1, \ldots, k \tag{24}$$

$$\sigma_j^2 \quad \sim \quad InvGam(1,1), \quad j = 1, \ldots, k. \tag{25}$$

The weights, means and variances are all updated using the Gibbs sampler but we use a Metropolis algorithm with a discrete uniform proposal for the allocation variables. Run lengths of 40000 will be used, discarding the first tenth as burn in. Behrens, Friel and Hurn (2012) considered this model when studying temperature placement for the tempered transition algorithm, finding that its expected log deviance curve had some interesting features. Figure 7 shows the estimated log deviance curves for $k = 1$ through to $k = 7$. The upper panel shows just the $k = 1$ curve as they are virtually indistinguishable due to the huge scale change as t approaches zero. However by restricting the y-axis, interesting differences can be seen for larger temperature values in the lower panel. Capturing these features efficiently for the numerical integration as k changes is the challenge. The top panel of Figure 8 shows how the adaptive scheme differs from the PF one for 200 rungs using $k = 1$ and $k = 3$. For $k = 1$, the adaptive scheme moves the points more densely towards zero where the expected log deviance decreases very rapidly. For $k = 3$, the expected log deviance also has a section of rapid change in the mid-range temperatures and the adaptive scheme attempts to capture this.

The massive gradient of the expected log deviance curves illustrates a potential pitfall of the adaptive temperature scheme. Figure 9 again shows a comparison between a rung of adaptive placement against the uncorrected PF scheme as the numbers of rungs increases. In the upper figure, the adaptive scheme is initialised at 10 PF spacings and in the lower at 20 PF spacings. In both cases by the time 200 rungs are being used, the adaptive version outperforms the PF in terms of the separation of the upper and lower bounds. But there is a dramatic difference at the start; the 10 rung initialisation takes far longer to catch up. At the end of the process, the two sets of $\{t_i\}$ are not very different (Figure 8, bottom panel). However the rate of change of the curve towards $t = 0$ is so great that the 10 rung version starts by filling in lots of temperatures close to zero, with the large differences in estimated gradients there meaning that our location procedure for additional points locates them very close together initially. There is at least an immediate diagnostic that the initial number of rungs is insufficient in that the corrected estimate lies outside the lower and upper bounds (Figure 9, upper panel); the changes in the gradient are so huge between $t = 0$ and the next few $\{t_i\}$ that pairwise differences do not form a good estimate of the second derivative (Equation (13)). This effect is the reason why we do not initialise the process using just the $t = 0$ and $t = 1$ points which are common to all schemes.

Using a 20 rung initialisation with 200 rungs in total, Figure 10 shows results for the mixture models with $k = 1$ to $k = 7$. The highest log evidence belongs to the $k = 3$ model, shortly followed by $k = 4, 5, 6$ and 7 in that order. The first set of non-overlapping discretisation bounds is between $k = 3$ and $k = 7$; the difference here between the corrected estimates is 3.38, with a corresponding Bayes factor of 29.48 for this



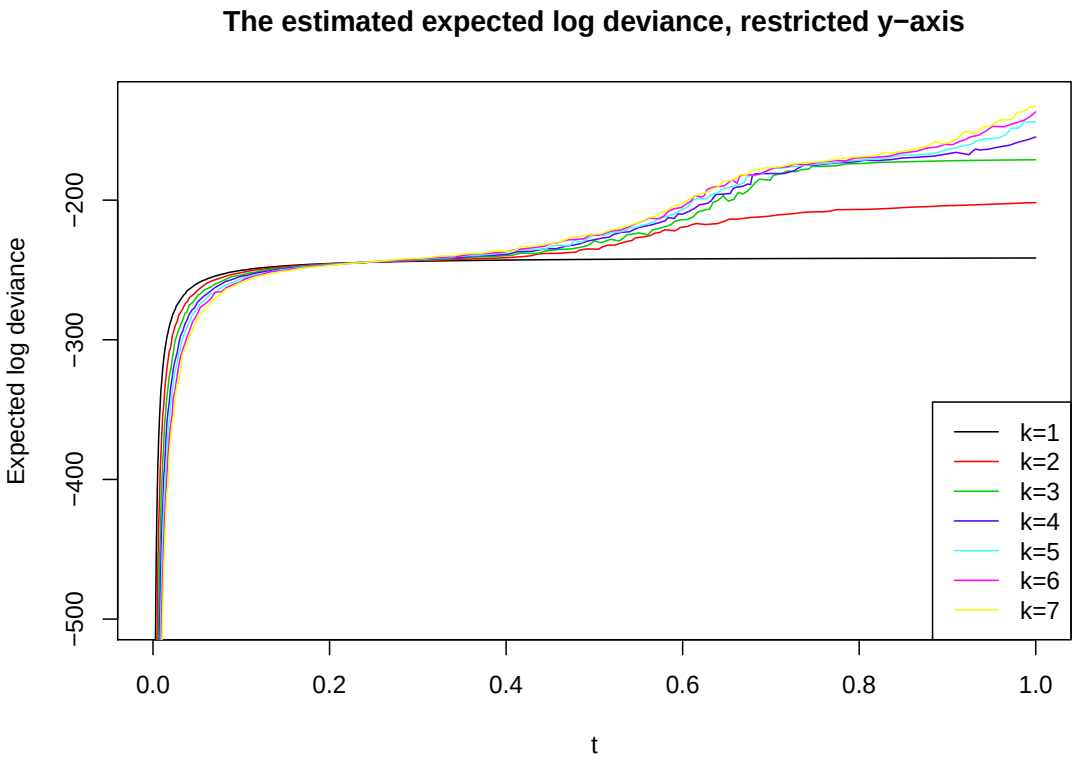
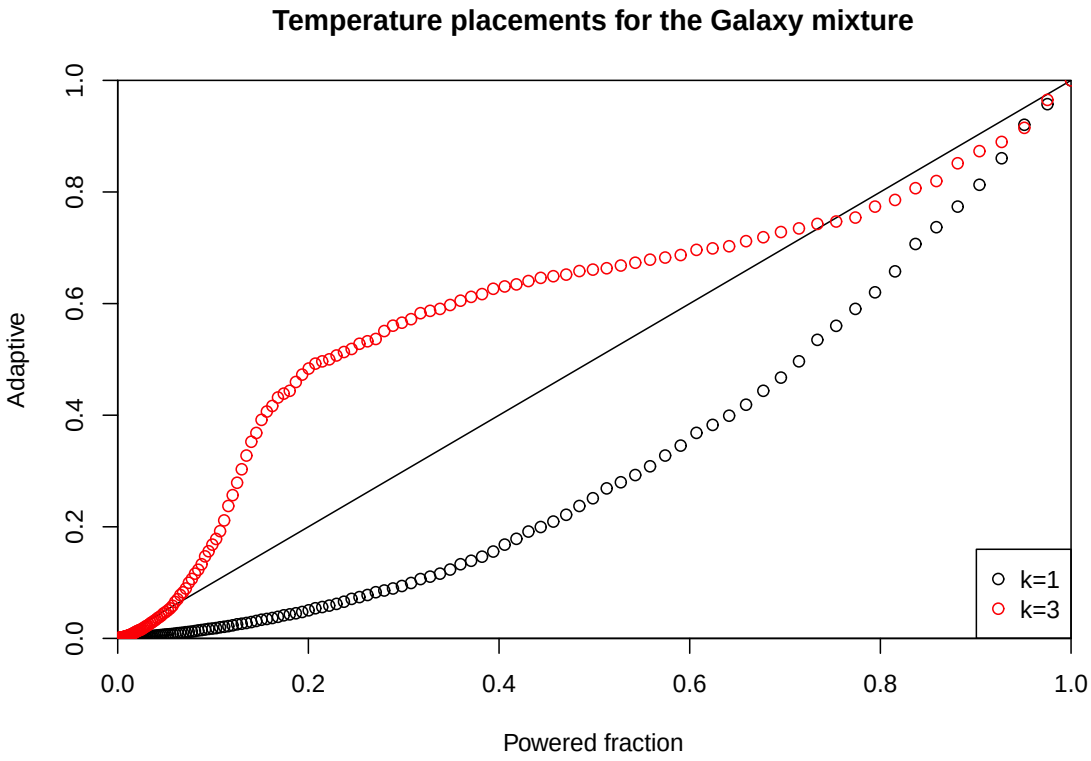


Figure 7: The expected log deviance curves for the Galaxy models using 200 rungs.



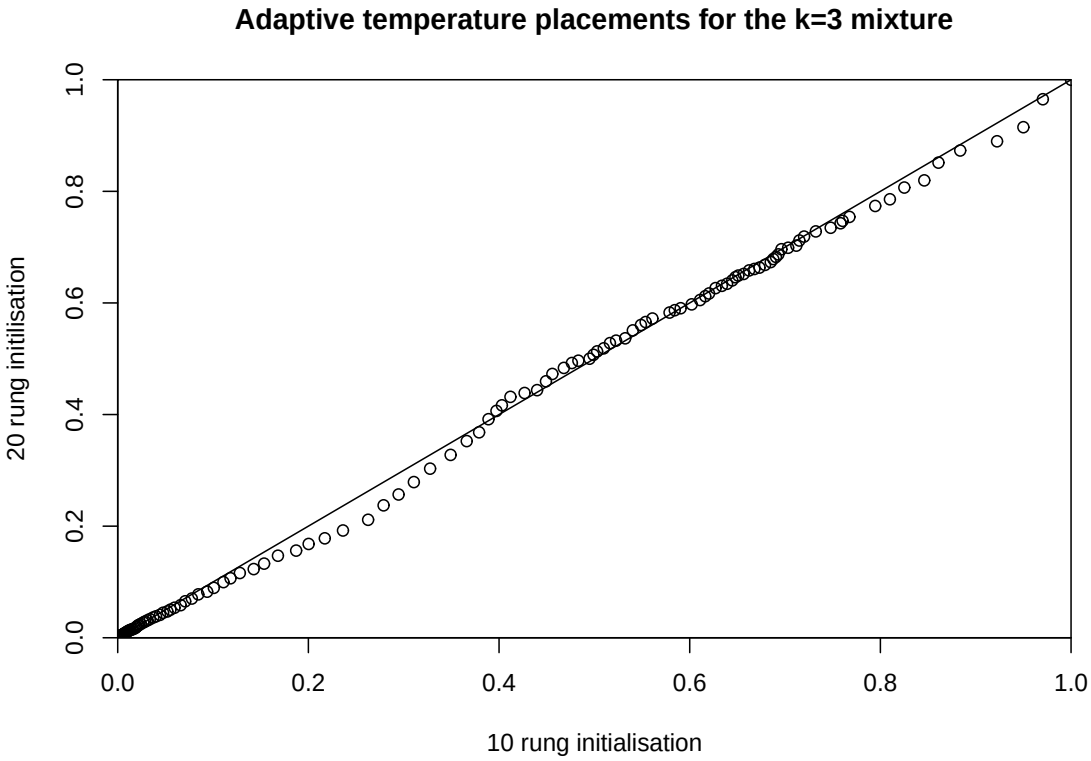
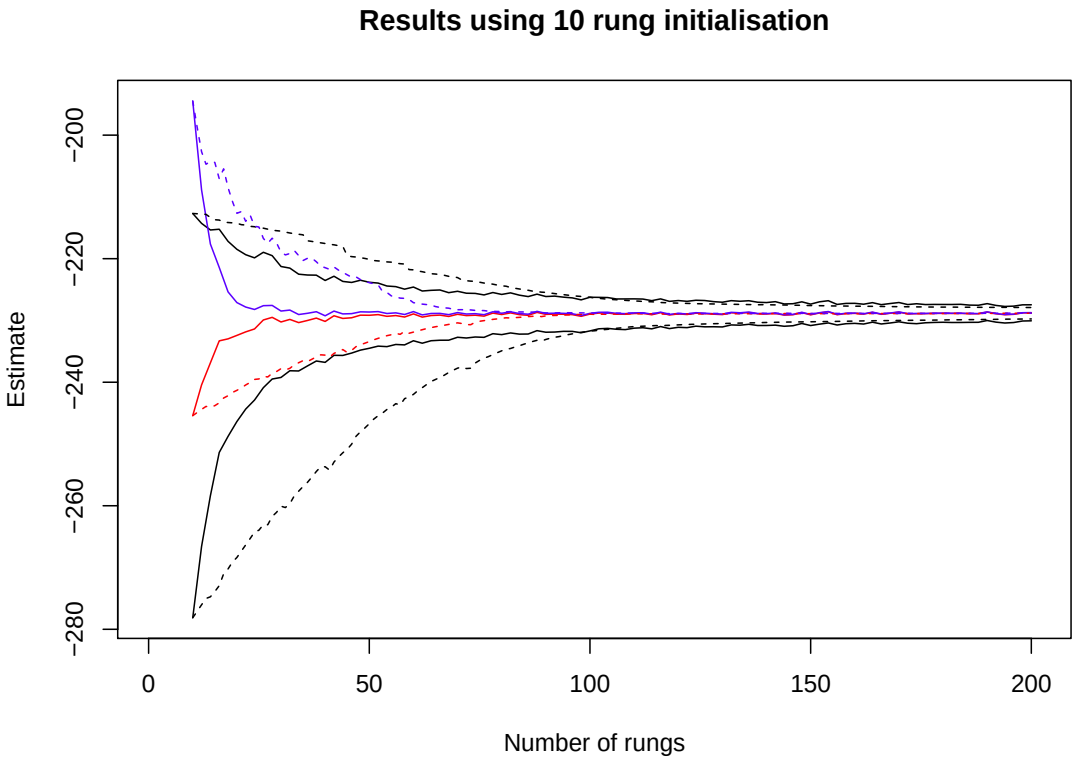


Figure 8: Top panel: The placement of the 200 rung schedules for the Galaxy data; the straight line indicates an exact match between PF and adaptive schemes. Bottom panel: The placements for $k = 3$ changing the initialisation from 10 to 20 rungs.



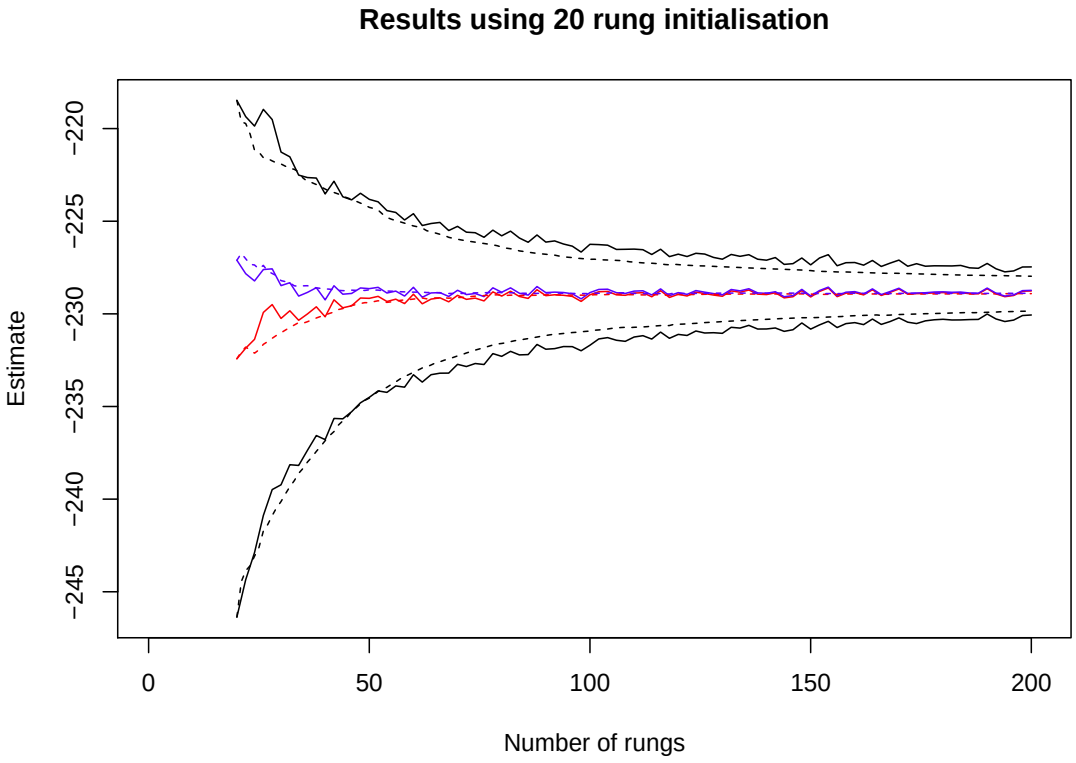


Figure 9: Upper and lower bounds (in black), uncorrected estimates (red), corrected estimates (blue) as the number of rungs increases for the $k = 3$ mixture model. Solid lines indicate PF spacing, dashed lines the adaptive schedule.

choice of priors. The adaptive scheme has the additional benefit that if these discretisation bounds are still considered too wide after using the anticipated maximum number of temperatures n, the process can simply be run on with additional temperatures placed by the same algorithm. The same cannot be said of the PF scheme where it is not immediately clear how to add additional temperatures.

4 Conclusions

This article has, we hope, illustrated the potential gains that can be made when estimating the evidence using power posteriors by correcting the numerical integration error and by adaptively choosing the temperature ladder. The methods that we have outlined come at virtually no extra computational cost, and we would therefore recommend that these are routinely used when implementing the power posterior approach. Given how effective the correction is but remembering that it does rely on good estimates of variance, it may be better to use a moderate size of n with long MCMC runs at each t_i, rather than dividing up the same total number of iterations into short runs with a large n.

What this article does not do is to give guidance as to how to allocate computational resources between the different temperatures. We have seen in our examples that the gradient and thus also the variance of $\mathbf{E}_{\theta|y,t} \log(p(y|\theta))$ is largest as $t \to 0$, suggesting that we should allocate more MCMC iterations here rather than as $t \to 0$ to get good estimates. On the other hand, when t is small, the power posteriors $p_t(\theta|y)$ will probably be easy to sample compared to when $t = 1$ and so we should also take into account the MCMC effective samples sizes. Neither the gradients nor the effective sample sizes can be known before sampling is carried out! There probably is some scope for an adaptation scheme here too, perhaps allocating some fraction of the total number of MCMC iterations evenly over the temperatures before allocating the remainder based on what we have learned in this initial phase. This point also reinforces our caveat: what we have addressed here is discretisation error, the bounds we give are (noisy) bounds on this error and not credible intervals in the usual sense.

In our examples, applying the correction term has effectively smoothed over the benefits of the adaptive scheme over the PF one (just working a little harder in the latter case). This numerical analysis trick is peculiar to this particular use of tempered distributions. In general we suspect though that the adaptation ideas developed here could find wider use in other tempered schemes described in the literature where numerical integration is not involved.

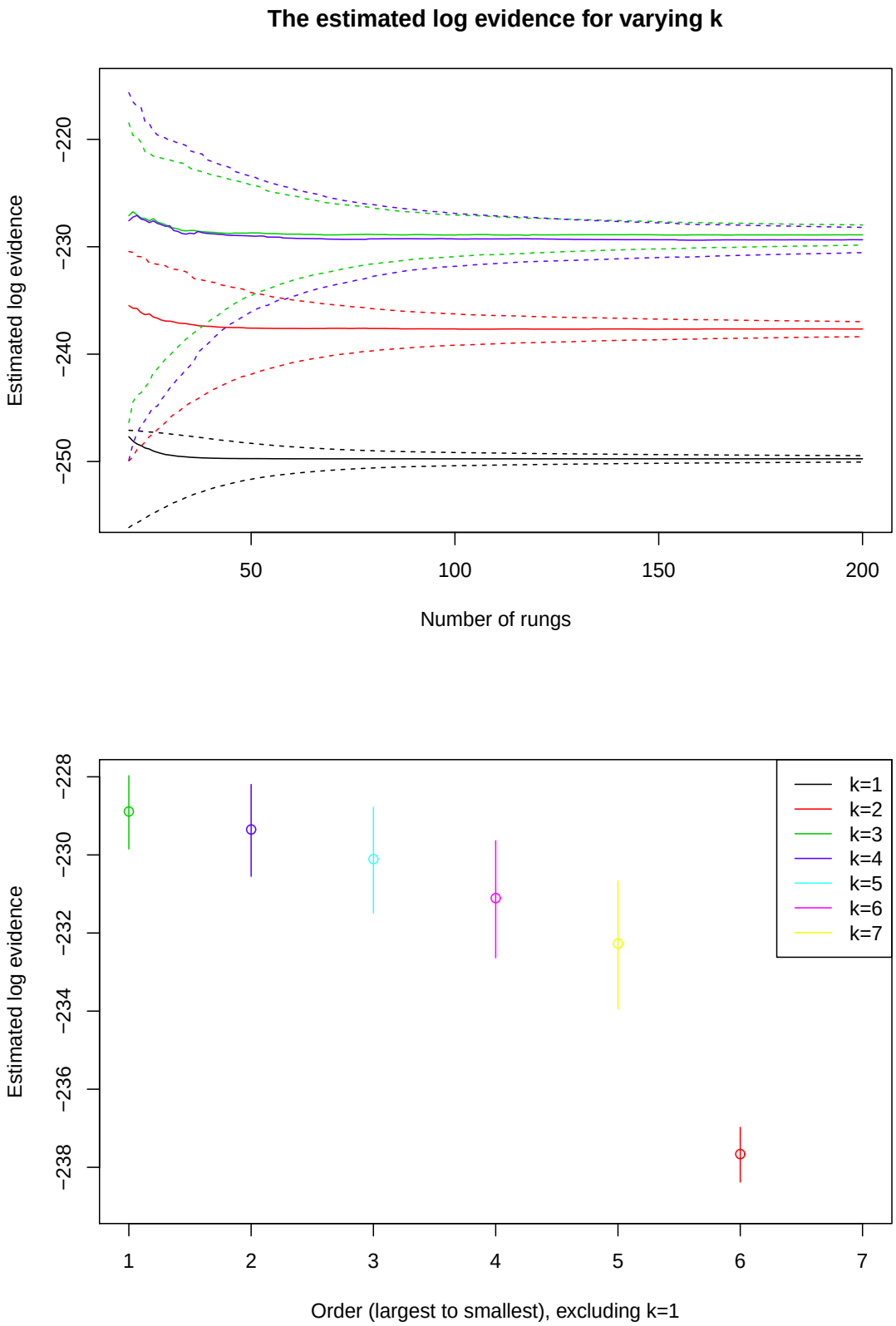


Figure 10: The estimated log evidence for the Galaxy data. Top panel: Dashed lines indicate upper and lower bounds, solid lines indicated the corrected estimate of the log evidence for $k = 1$ through to $k = 4$. Bottom panel: vertical bars indicate the width of the lower and upper bound, circles the corrected estimates, $k = 2$ through to $k = 7$.

Acknowledgements: Nial Friel's research was supported by a Science Foundation Ireland Research Frontiers Program grant, 09/RFP/MTH2199. Jason Wyse's research was supported through the STATICA project, a Principal Investigator program of Science Foundation Ireland, 08/IN.1/I1879.

References

[1] K. Atkinson and W. Han. *Elementary numerical analysis*. John Wiley and sons, third edition, 2004.

[2] G. Behrens, N. Friel, and M. Hurn. Tuning tempered transitions. *Statistics and Computing*, 22(1):65–78, 2012.

[3] B. Calderhead and M. Girolami. Estimating Bayes factors via thermodynamic integration and population mcmc. *Computational Statistics & Data Analysis*, 53(12):4028–4045, 2009.

[4] S. Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, 1995.

[5] N. Friel and A.N. Pettitt. Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society B*, 70(3):589–607, 2008.

[6] N. Friel and J. Wyse. Estimating the evidence - a review. *Statistica Neerlandica*, 66(3):288–308, 2012.

[7] P.J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[8] N. Lartillot and H. Philippe. Computing Bayes factors using thermodynamic integration. *Systematic Biology*, 55(2):195–207, 2006.

[9] G. Lefebvre, R.J. Steele, and A.C. Vandal. A path sampling identity for computing the Kullback-Leibler and J-divergences. *Computational Statistics and Data Analysis*, 54(7):1719–1731, 2010.

[10] X.L. Meng and W.H. Wong. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, 6(4):831–860, 1996.

[11] R.M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

[12] S. Richardson and P.J. Green. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society B*, 59(4):731–792, 1997.

[13] J. Skilling. Nested sampling for general Bayesian computation. *Bayesian Analysis*, 1(4):833–860, 2006.

[14] J.W. Smith, J.E. Everhart, W.C. Dickson, W.C. Knowler, and R.S. Johannes. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 261. American Medical Informatics Association, 1988.

[15] L. Tierney and J.B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986.

[16] E. Williams. *Regression Analysis*. Wiley, Chichester, 1959.